ENV 790.30 - Time Series Analysis for Energy Data | Spring 2023

Assignment 2 - Due date 02/03/23

Yuxiang Ren

Submission Instructions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github.

Once you have the file open on your local machine the first thing you will do is rename the file such that it includes your first and last name (e.g., "LuanaLima_TSA_A02_Sp23.Rmd"). Then change "Student Name" on line 4 with your name.

Then you will start working through the assignment by **creating code and output** that answer each question. Be sure to use this assignment document. Your report should contain the answer to each question and any plots/tables you obtained (when applicable).

When you have completed the assignment, **Knit** the text and code into a single PDF file. Submit this pdf using Sakai.

R packages

R packages needed for this assignment:"forecast","tseries", and "dplyr". Install these packages, if you haven't done yet. Do not forget to load them before running your script, since they are NOT default packages.\

```
#Load/install required package here
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(tseries)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(xlsx)
```

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source on our **Data** folder. The data comes from the US Energy Information and Administration and corresponds to the December 2022 Monthly Energy Review. The spreadsheet is ready to be used. You will also find a *.csv* version of the data "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source-Edit.csv". You may use the function *read.table*() to import the *.csv* data in R. Or refer to the file "M2_ImportingData_CSV_XLSX.Rmd" in our Lessons folder for functions that are better suited for importing the *.xlsx*.

```
#Importing data set

A02_rawdata <- read.xlsx(file="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.x
read_col_names <- read.xlsx(file="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sourc

colnames(A02_rawdata) <- read_col_names
head(A02_rawdata)
```

```
##        Month Wood Energy Production Biofuels Production
## 1 1973-01-01                129.630       Not Available
## 2 1973-02-01                117.194       Not Available
## 3 1973-03-01                129.763       Not Available
## 4 1973-04-01                125.462       Not Available
## 5 1973-05-01                129.624       Not Available
## 6 1973-06-01                125.435       Not Available
##   Total Biomass Energy Production Total Renewable Energy Production
## 1                         129.787                            403.981
## 2                         117.338                            360.900
## 3                         129.938                            400.161
## 4                         125.636                            380.470
## 5                         129.834                            392.141
## 6                         125.611                            377.232
##   Hydroelectric Power Consumption Geothermal Energy Consumption
## 1                         272.703                         1.491
## 2                         242.199                         1.363
## 3                         268.810                         1.412
## 4                         253.185                         1.649
## 5                         260.770                         1.537
## 6                         249.859                         1.763
##   Solar Energy Consumption Wind Energy Consumption Wood Energy Consumption
## 1           Not Available          Not Available                129.630
## 2           Not Available          Not Available                117.194
## 3           Not Available          Not Available                129.763
## 4           Not Available          Not Available                125.462
## 5           Not Available          Not Available                129.624
## 6           Not Available          Not Available                125.435
##   Waste Energy Consumption Biofuels Consumption
## 1                    0.157        Not Available
## 2                    0.144        Not Available
```

```
## 3                    0.176        Not Available
## 4                    0.174        Not Available
## 5                    0.210        Not Available
## 6                    0.176        Not Available
##   Total Biomass Energy Consumption Total Renewable Energy Consumption
## 1                          129.787                             403.981
## 2                          117.338                             360.900
## 3                          129.938                             400.161
## 4                          125.636                             380.470
## 5                          129.834                             392.141
## 6                          125.611                             377.232
```

Question 1

You will work only with the following columns: Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption. Create a data frame structure with these three time series only. Use the command head() to verify your data.

```
#Q1
df_biomP_renewP_hydroC <- data.frame("BiomP"=A02_rawdata$`Total Biomass Energy Production`,
                                     "RenewP"=A02_rawdata$`Total Renewable Energy Production`,
                                     "HydroC"=A02_rawdata$`Hydroelectric Power Consumption`)

head(df_biomP_renewP_hydroC)
```

```
##     BiomP  RenewP  HydroC
## 1 129.787 403.981 272.703
## 2 117.338 360.900 242.199
## 3 129.938 400.161 268.810
## 4 125.636 380.470 253.185
## 5 129.834 392.141 260.770
## 6 125.611 377.232 249.859
```

```
#with time
df_date_brh <- cbind(A02_rawdata$Month,df_biomP_renewP_hydroC)
names(df_date_brh) <- c("Time", "BiomP", "RenewP", "HydroC")
```

Question 2

Transform your data frame in a time series object and specify the starting point and frequency of the time series using the function ts().

```
ts_df_biomP_renewP_hydroC <- ts(df_biomP_renewP_hydroC[,1:3], frequency = 12, start = c(1973, 1))
```

Question 3

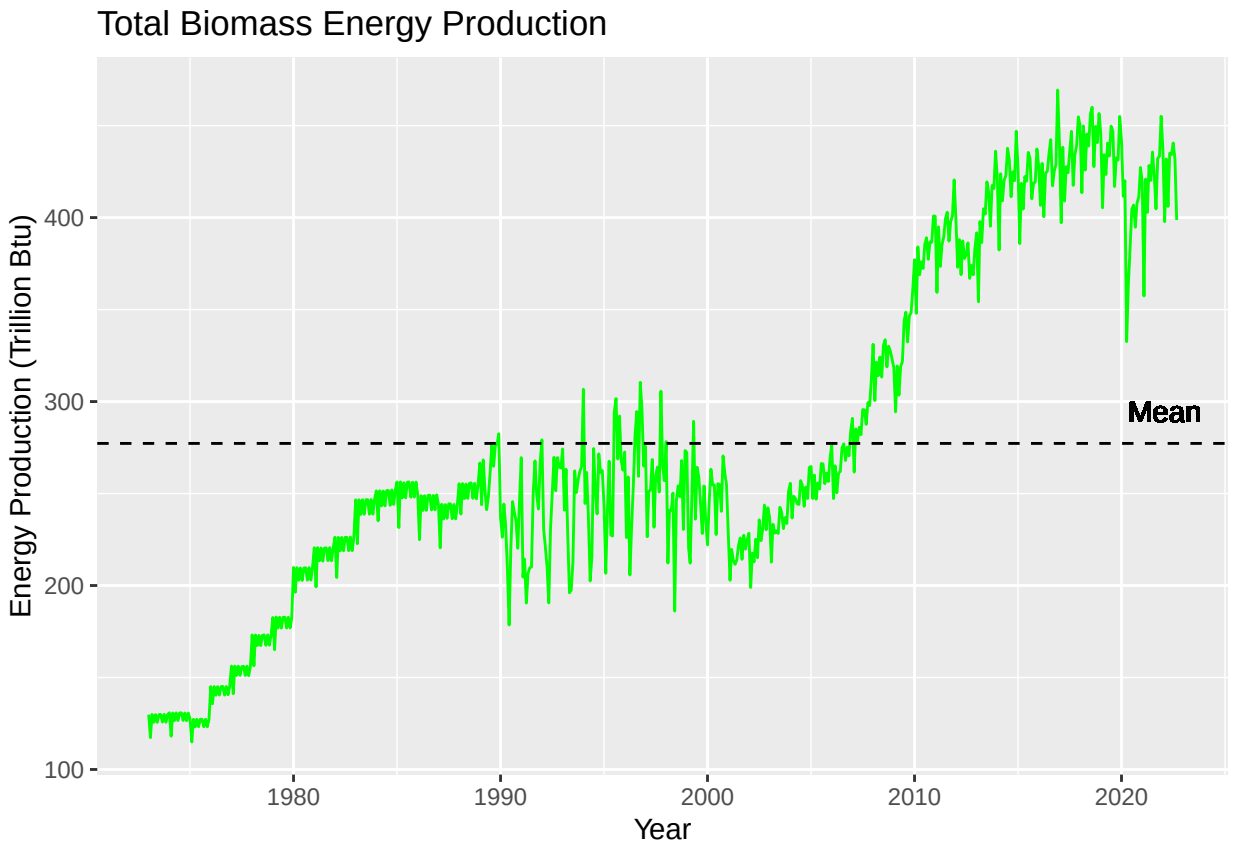
Compute mean and standard deviation for these three series.

```r
#Total biomass energy production
mean_biomP <- mean(df_biomP_renewP_hydroC$BiomP)
sd_biomP <- sd(df_biomP_renewP_hydroC$BioP)

#Total renewable energy production
mean_renewP <- mean(df_biomP_renewP_hydroC$RenewP)
sd_renewP <- sd(df_biomP_renewP_hydroC$RenewP)

#Hydroelectric power consumption
mean_hydroC <- mean(df_biomP_renewP_hydroC$HydroC)
sd_hydroC <- sd(df_biomP_renewP_hydroC$HydroC)
```

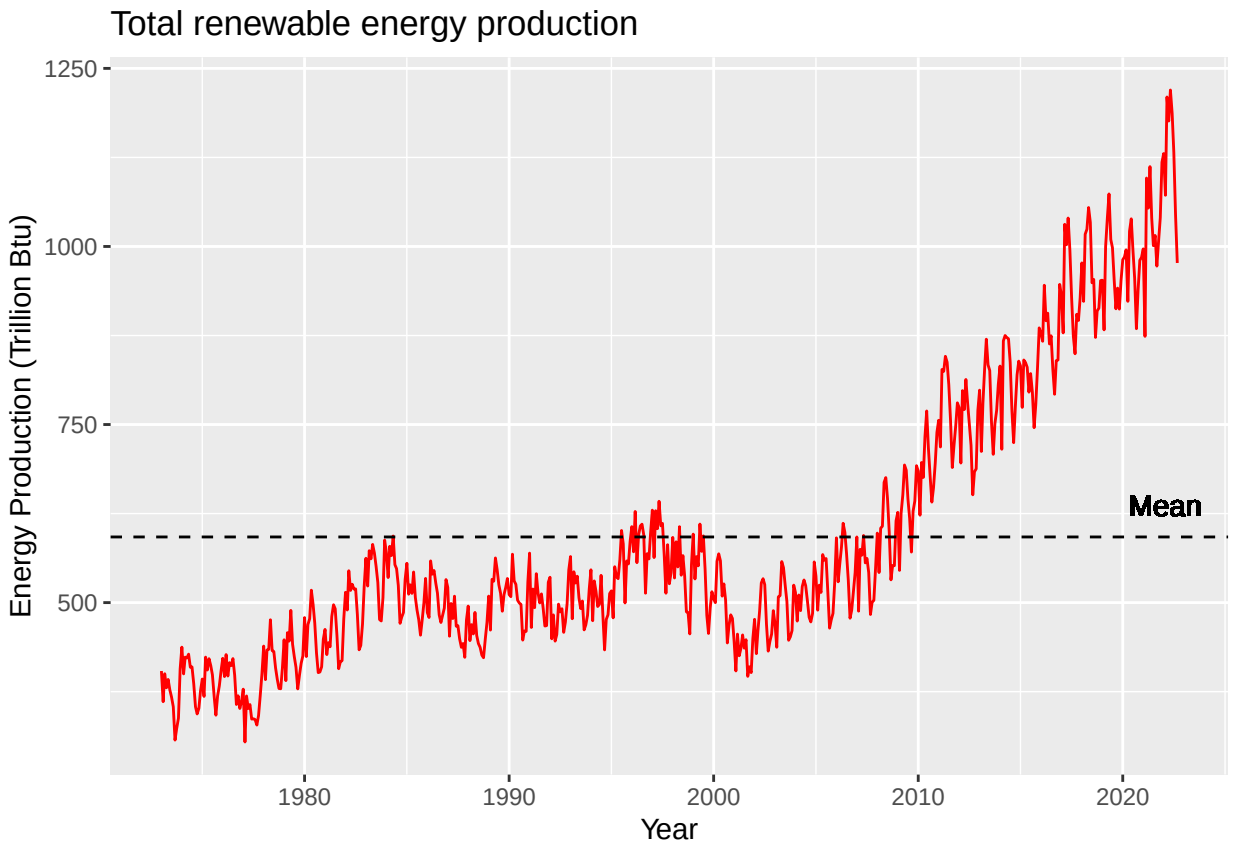
Question 4

Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

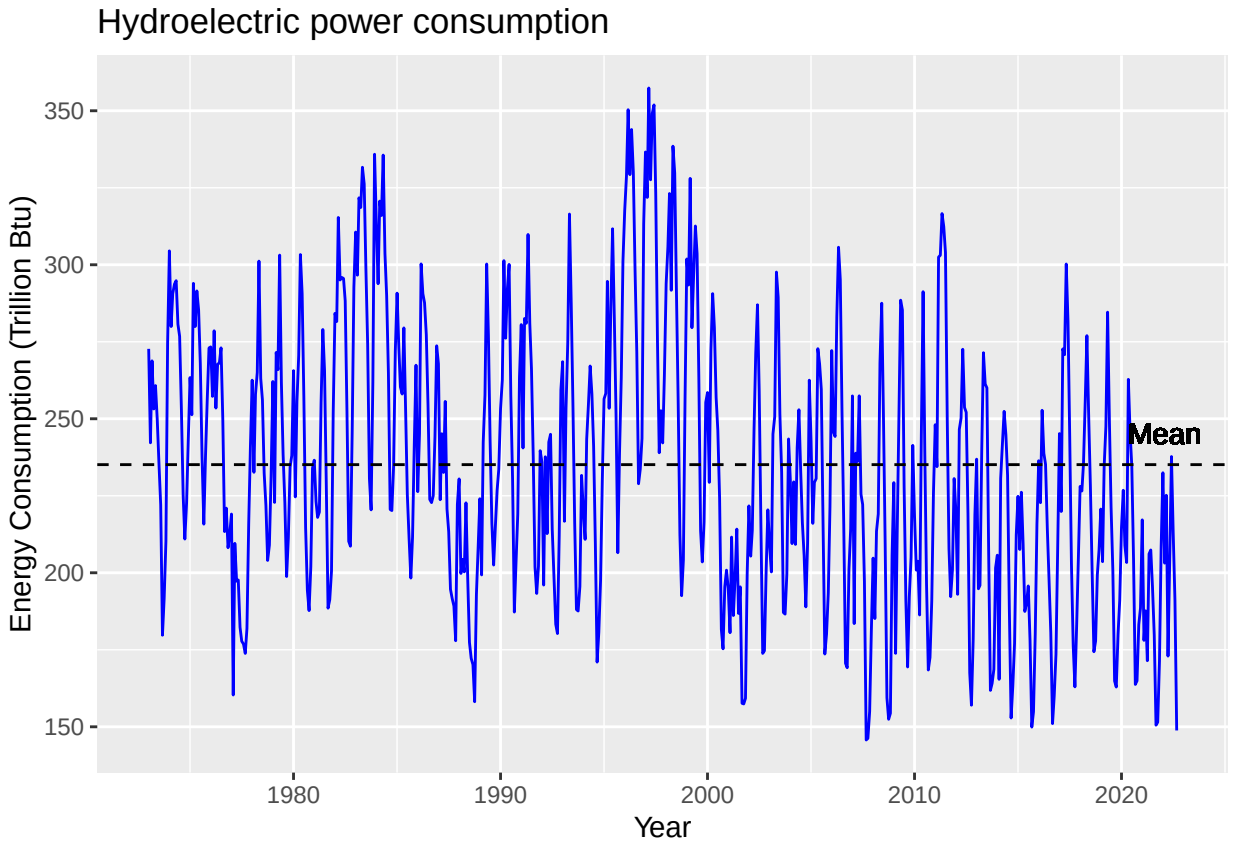
```r
#date for label
date <- df_date_brh[590,1]

#Total biomass energy production
plot_bioP <- ggplot(df_date_brh, aes(x=Time, y=BiomP)) +
            geom_line(color="Green")+
  geom_hline(aes(yintercept=mean_biomP), colour= "black", linetype="dashed")+
  geom_text(aes(date, mean_biomP, label = "Mean", vjust= -1))+
  ggtitle("Total Biomass Energy Production")+ xlab("Year")+
  ylab("Energy Production (Trillion Btu)")
plot_bioP
```




```
#Total renewable energy production
plot_renewP <- ggplot(df_date_brh, aes(x=Time, y=RenewP)) +
            geom_line(color="red")+
  geom_hline(aes(yintercept=mean_renewP), colour= "black", linetype="dashed")+
  geom_text(aes(date, mean_renewP, label = "Mean", vjust= -1))+
  ggtitle("Total renewable energy production")+ xlab("Year")+
  ylab("Energy Production (Trillion Btu)")
plot_renewP
```




```
#Hydroelectric power consumption
plot_hydroC <- ggplot(df_date_brh, aes(x=Time, y=HydroC)) +
            geom_line(color="blue")+
  geom_hline(aes(yintercept=mean_hydroC), colour= "black", linetype="dashed")+
  geom_text(aes(date, mean_hydroC, label = "Mean", vjust= -1))+
  ggtitle("Hydroelectric power consumption")+ xlab("Year")+
  ylab("Energy Consumption (Trillion Btu)")
plot_hydroC
```



Answer:

(1)The overall trend for total biomass energy production is increasing. The obvious upward changes are between 1975 and 1990, and from 2000 to 2017. It can also be observed that at the beginning of 2020, biomass energy production suddenly dropped sharply and has gradually increased in recent years. The figure shows seasonal trend. (2)The overall trend for total renewable energy production is increasing. The increasing trend has become noticeable since 2000. The figure shows strong seasonality. (3)The overall change trend of the total consumption of hydroelectric power consumption is a slight decline. It can be seen that most of the monthly energy consumption before 2000 is greater than that after 2000.The figure shows strong seasonality.

Question 5

Compute the correlation between these three series. Are they significantly correlated? Explain your answer.

```
#correlation between three series
cor_all <- cor(df_biomP_renewP_hydroC[, c(1,2,3)])
cor_all
```

```
##             BiomP      RenewP      HydroC
## BiomP   1.0000000  0.91859411 -0.29982013
## RenewP  0.9185941  1.00000000 -0.09958758
## HydroC -0.2998201 -0.09958758  1.00000000
```

```r
# biomP and renewP
cor_br <- cor.test(df_biomP_renewP_hydroC$BiomP,df_biomP_renewP_hydroC$RenewP)
cor_br
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_biomP_renewP_hydroC$BiomP and df_biomP_renewP_hydroC$RenewP
## t = 56.697, df = 595, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9050636 0.9302668
## sample estimates:
##       cor
## 0.9185941
```

```r
# biomP and hydroC
cor_bh <- cor.test(df_biomP_renewP_hydroC$BiomP,df_biomP_renewP_hydroC$HydroC)
cor_bh
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_biomP_renewP_hydroC$BiomP and df_biomP_renewP_hydroC$HydroC
## t = -7.6661, df = 595, p-value = 7.256e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3711363 -0.2249878
## sample estimates:
##        cor
## -0.2998201
```

```r
# renewP and hydroC
cor_rh <- cor.test(df_biomP_renewP_hydroC$RenewP,df_biomP_renewP_hydroC$HydroC)
cor_rh
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_biomP_renewP_hydroC$RenewP and df_biomP_renewP_hydroC$HydroC
## t = -2.4413, df = 595, p-value = 0.01492
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17840723 -0.01949801
## sample estimates:
##         cor
## -0.09958758
```

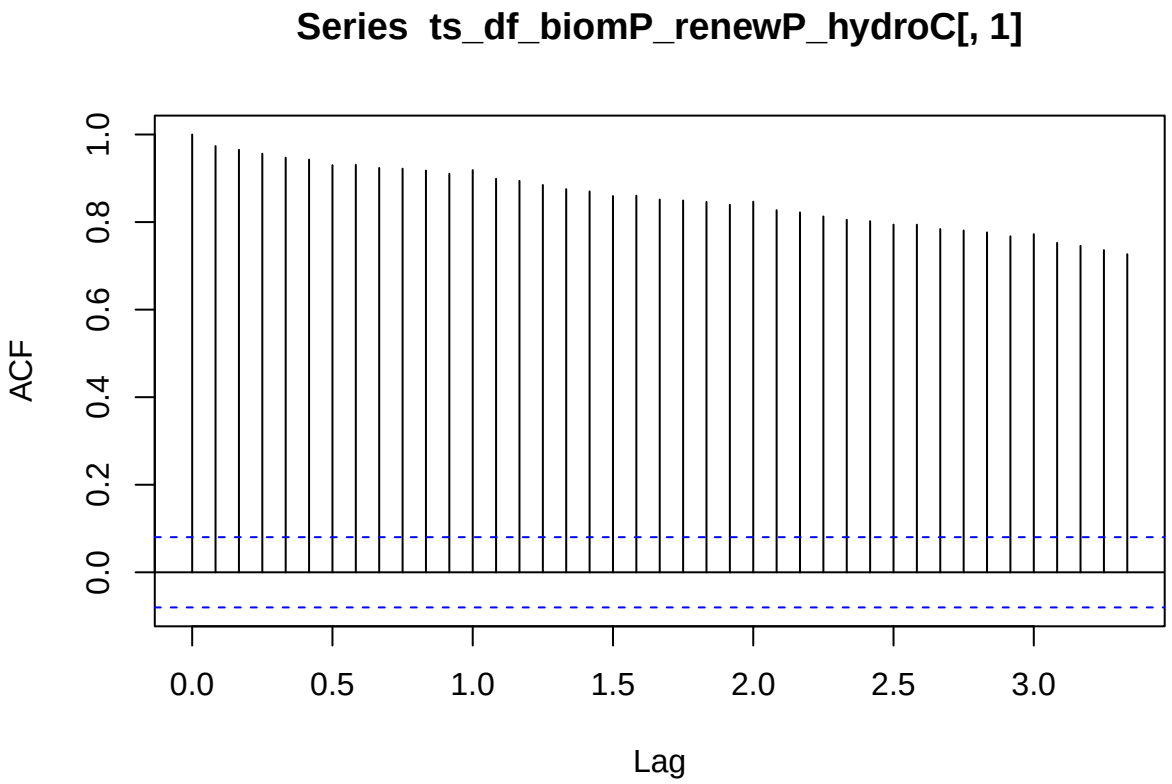
Answer: (1) Biomass energy production and renewable energy production have a strong positive correlation, which coefficient is 0.919. Additionally, the p-value is less than 0.05, the result is significant. (2) Biomass energy production and hydroelectric power consumption have a low negative correlation. The correlation coefficient of these two variables is -0.300. And because

the p-value is less than 0.05, the result is significant. (3) Renewable energy production and hydroelectric power consumption have negligible or weak correlation due to the coefficient being -0.100. The p-value of these two variables is less than 0.05. The result is significant.

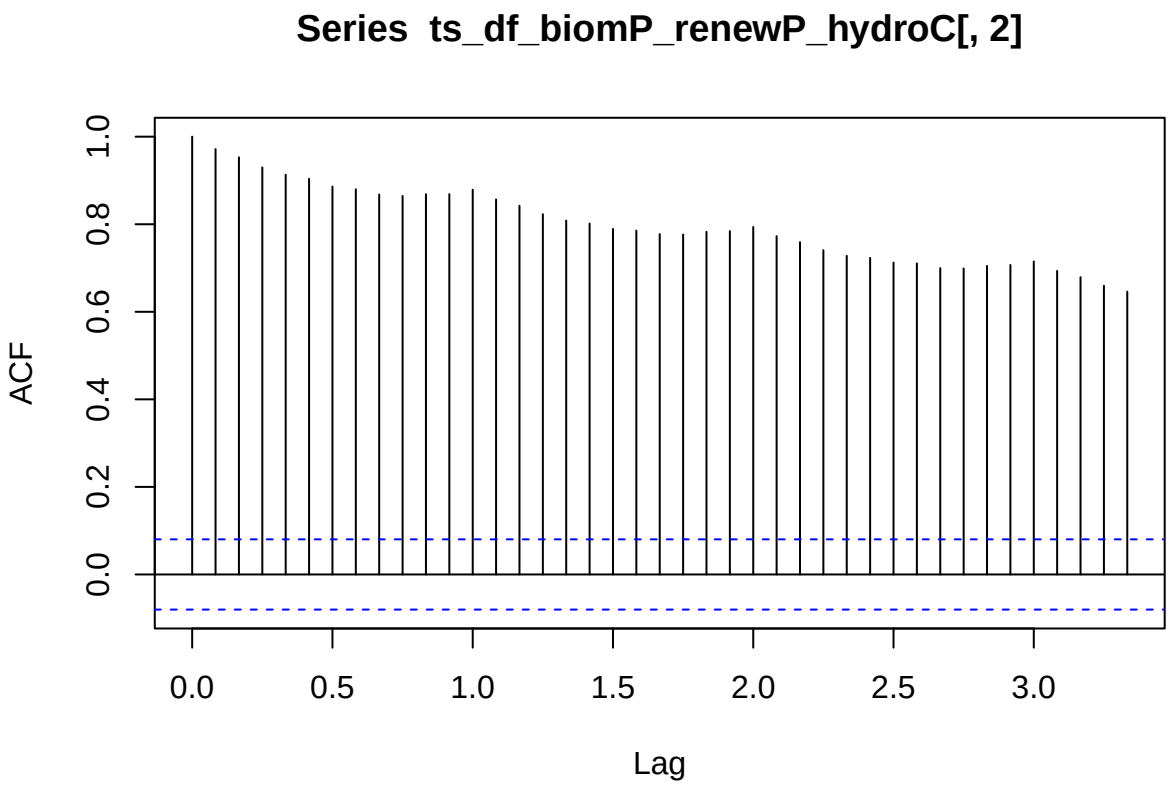
Question 6

Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

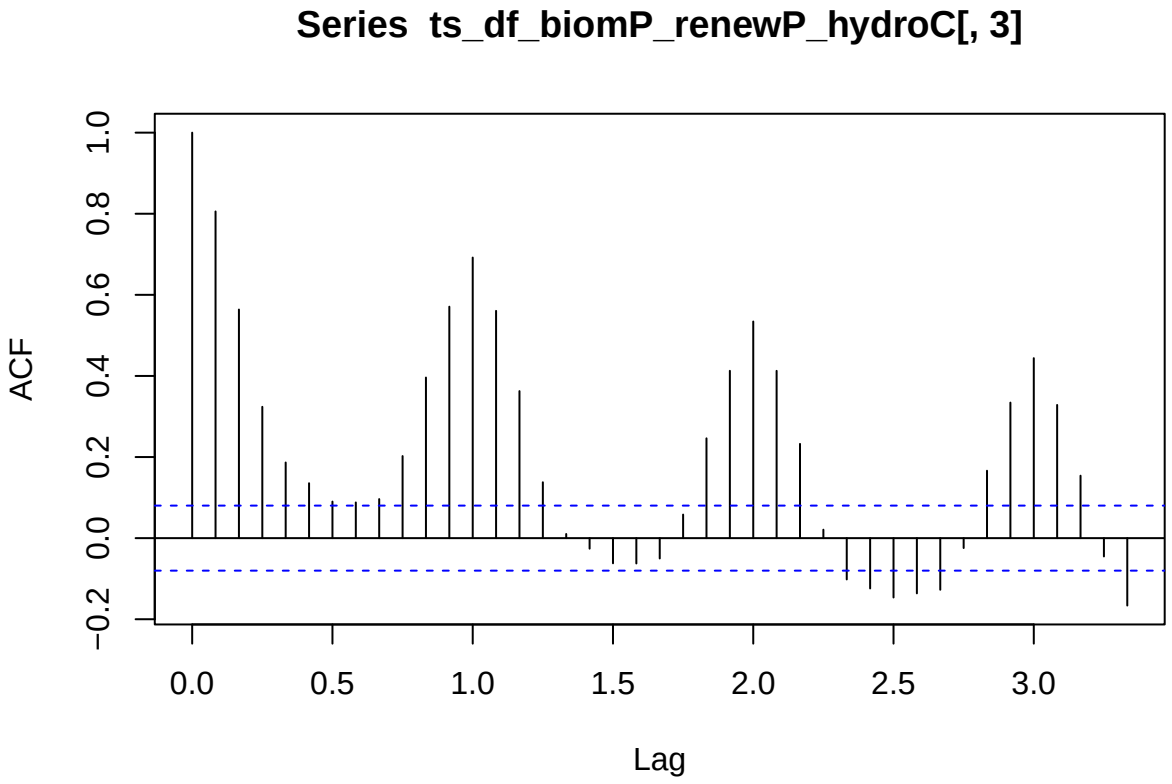
```
biomP_acf=acf(ts_df_biomP_renewP_hydroC[,1],lag.max=40, type="correlation", plot=TRUE)
```




```
renewP_acf=acf(ts_df_biomP_renewP_hydroC[,2],lag.max=40, type="correlation", plot=TRUE)
```




```
hydroC_acf=acf(ts_df_biomP_renewP_hydroC[,3],lag.max=40, type="correlation", plot=TRUE)
```



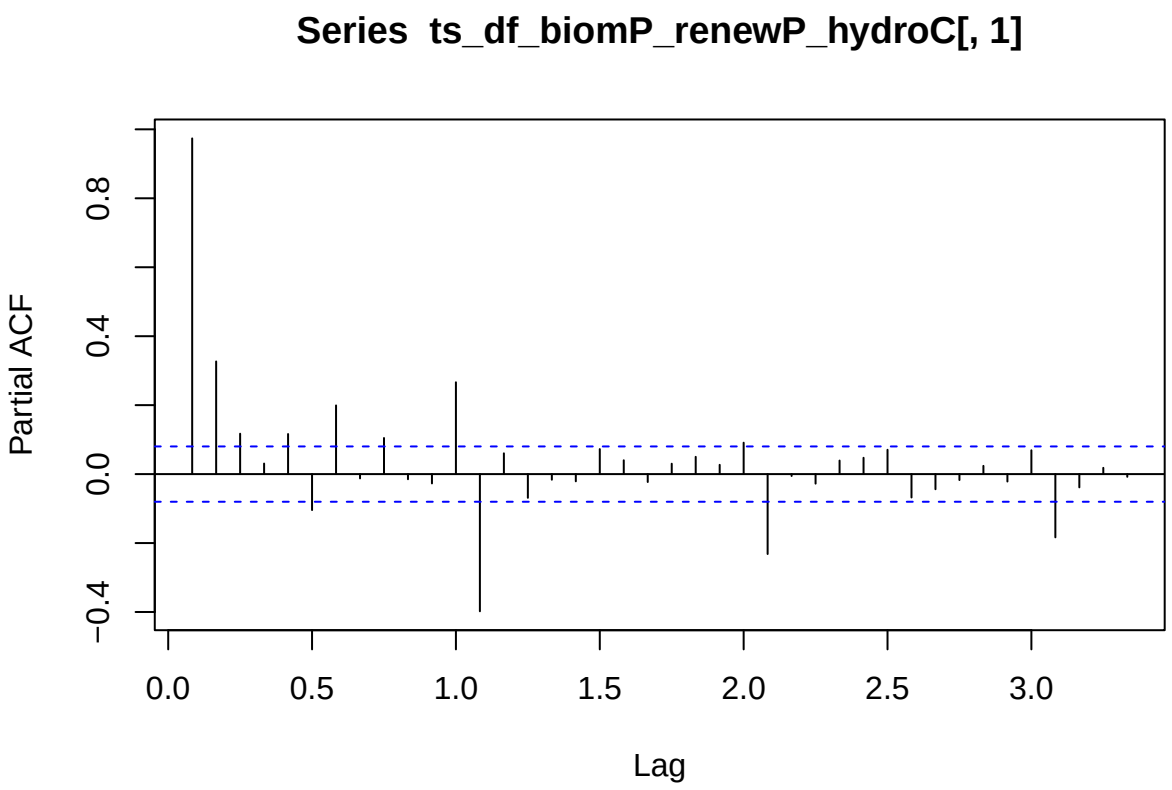
Answer: Both biomass and renewable energy production have a high degree of autocorrelation

between themselves and lagged versions. Although both biomass and renewable's ACF are declining with lags, all lags are significant. However, some lags' value in the ACF of hydroelectric power consumption is insignificant. The change in the autocorrelation value of lag shows that the data might have some seasonality.

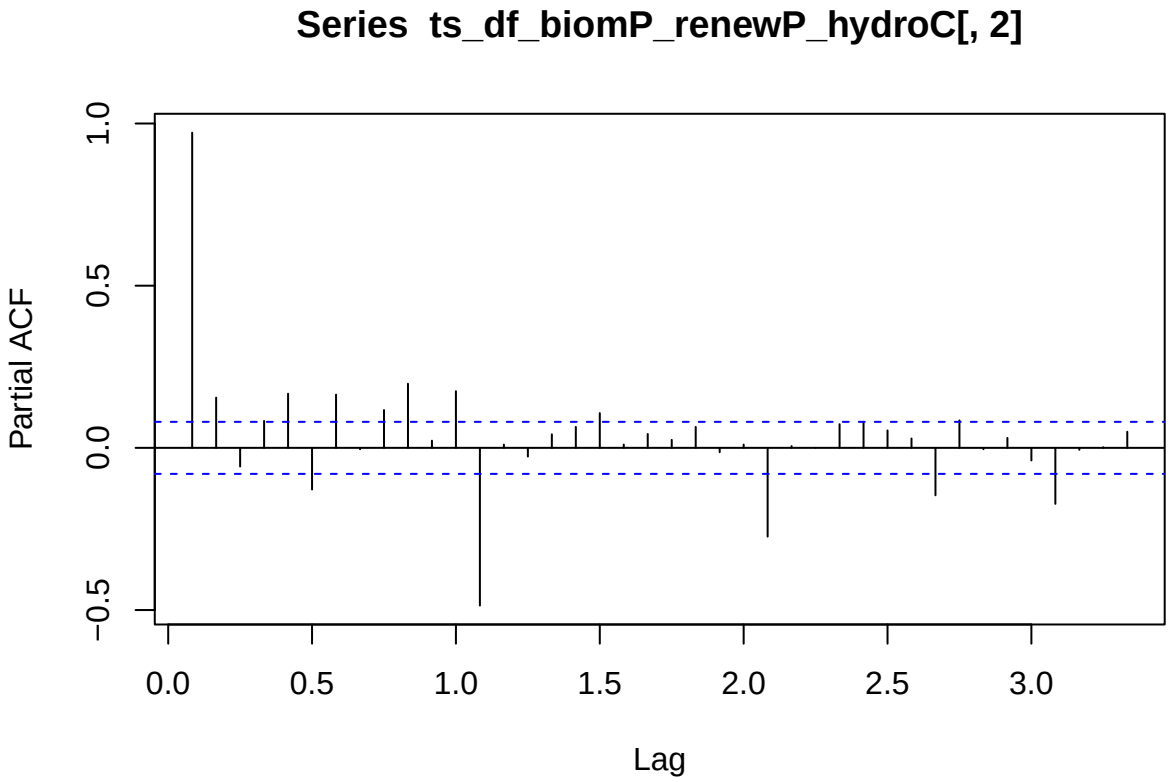
Question 7

Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

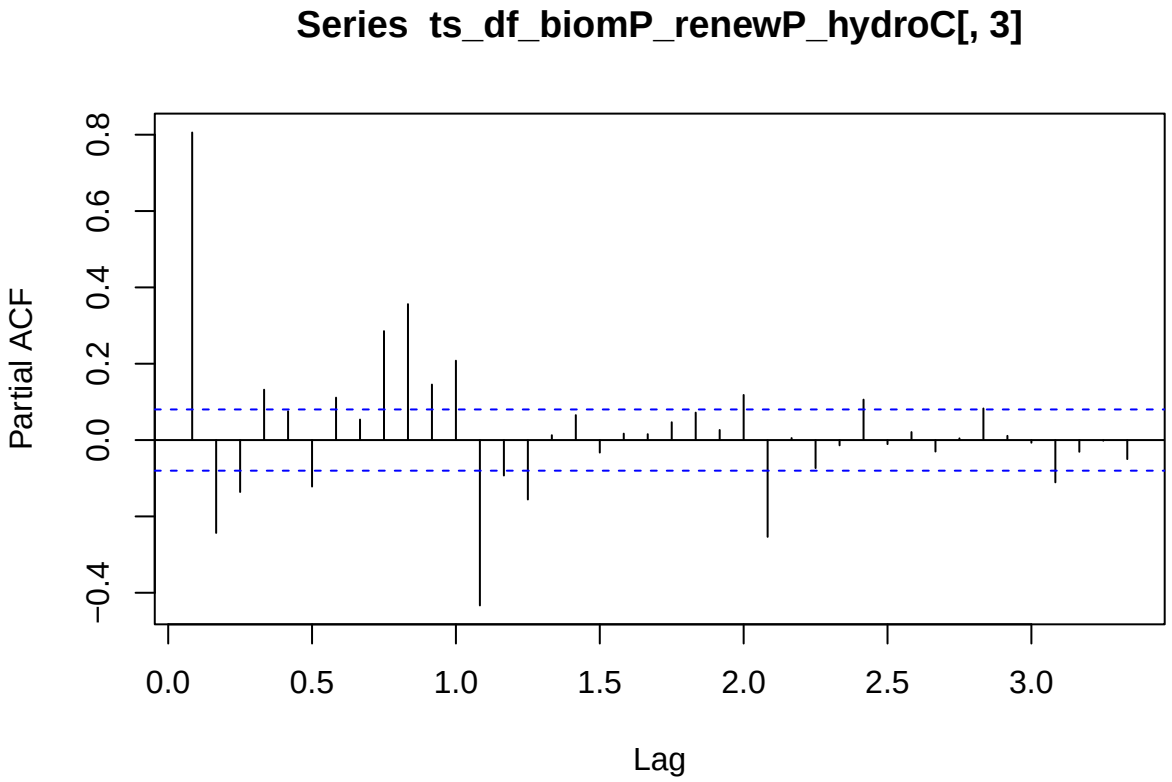
```
biomP_pacf=pacf(ts_df_biomP_renewP_hydroC[,1],lag.max=40, plot=TRUE)
```




```
renewP_pacf=pacf(ts_df_biomP_renewP_hydroC[,2],lag.max=40, plot=TRUE)
```




```
hydroC_pacf=pacf(ts_df_biomP_renewP_hydroC[,3],lag.max=40, plot=TRUE)
```



Answer: Most lags' value drops and becomes insignificant. Meanwhile, more significant negative values emerge.